A Unified Framework for Risk-sensitive Markov Decision Processes with Finite State and Action Spaces

Yun Shen
Technical University of Berlin
Berlin, 10587, Germany
yun@cs.tu-berlin.de

Steffen Grünewälder
University College London
London WC1E 6BT, UK
s.grunewalder@cs.ucl.ac.uk

Klaus Obermayer
Technical University of Berlin
Berlin, 10587, Germany
oby@cs.tu-berlin.de

November 1, 2011

Abstract

We introduce a unified framework to incorporate risk in Markov decision processes (MDPs), via prospect maps, which generalize the idea of coherent/convex risk measures in mathematical finance. Most of the existing risk-sensitive approaches in various literature concerning with decision-making problems are contained in the framework as special instances. Within the framework, we solve the optimal control problems according to two criteria, the newly invented temporal discounted criterion, which generalizes the conventional discount scheme, and the average criterion, by value iteration algorithms under different assumptions. Two online algorithms are proposed to solve the optimal controls problem when the exact MDP is unknown and has to be estimated during optimization.

1 Introduction

In many applications of decision-making problems modeled by *Markov decision processes* (MDPs), it is reasonable to incorporate some measure of risk to rule out policies that achieve a high expected reward at the cost of risky and error prone actions. If we think for example of an expensive manufacturing machine that has two running modes: one where the machine runs at peak level and produces the maximum number of products for most of the time at the cost of a high chance for a serious damage and one where the machine runs slightly slower to avoid damage. Most companies would agree that the second option is more reasonable. Yet, if the company would make decision with the help of the classical MDPs, it would pick option one and go for the risky strategy.

Most of the decision-making models like MDPs, are consisted with two descriptions of some mechanism of environments, immediate *outcomes* (rewards or costs) at one state by performing one action, and transitions, the transition probability between states with some actions. Both descriptions are *objective* in the sense that both outcome and transition probability can be estimated by repeating experiencing the environment sufficient many times. The "risk" depends, however, on the *subjective* perception of the agent, since different agents might have different *risk-preferences* facing the same environment. For instance, \$100 is more valuable for the poor than for the rich.

Behavioral experiments [21] show that people tend to *overreact* to small probability events, but *underreact* to medium and large probabilities.

Due to the apparent usefulness of risk-sensitive objectives, the topic is of major importance in finance and economics. In economics, the *utility function* is widely used to model the subjective perception of rewards. The renowned *prospect theory* (PT) [21] introduces the *probability weighting function* to model the subjective perception of probabilities. PT can be merely used to model single decision problem, whereas in MDP a sequence of decisions have to be made. In mathematical finance, Ruszczyński (2010) [28] applies *coherent/convex risk measures* (CRMs) [2, 11] to incorporate risk in a sequential decision-making structure. However, there are two major drawbacks in their work: 1) he assumes that the risk measures must be *coherent* or *convex*, which is not true for some of the most important instances of risk measures, and 2) he discusses merely the finite-stage or discounted risk problem for coherent risk measures. The theory of discounted and average risk for arbitrary measures as in the classical MDP have not been considered yet.

In the community of MDPs (mainly operations research and control theory), despite the apparent usefulness of risk-sensitive measures, few works in MDPs address the issue, since many risk-sensitive objectives cannot be optimized efficiently. The mean-variance trade-off is a popular risk criterion, where variance takes the part of the risk measure as it penalizes highly varying return. However, this objective is difficult to optimize, especially when a discount factor is included [10]. Recently in [23] the problem even for finite-horizon MDP is proved to be NP-hard. Another popular measure is to apply the exponential utility function. Although an efficient solution (see e.g. [5]) exists for average infinite-horizon MDP, it is proved in [7] that the objective for discounted MDP is difficult and the optimal policy might not be stationary.

The question is now if all the risk-sensitive objectives are difficult to optimize for MDPs or if measures like the mean-variance trade-off are just not the "right" measure for MDPs. Inspired by the discovery in mathematical finance and economics, our intuition is therefore to adapt the CRM theory to the MDP structure, where two concerns must be balanced: 1) the axioms should be as general as possible to be able to model all kinds of risk-preferences including mixed risk-preference, and 2) the underlying optimization problem can be solved by a computationally feasible algorithm.

The main contributions of this paper are: 1) To incorporate risk into MDPs, we set up a general framework via *prospect maps*, which is a generalization of the CRMs. The framework contains most of the existing risk-sensitive approaches in economics, mathematical finance and optimal control theory as special cases (cf. Sec. 5). 2) Within the framework, we define a novel temporal discount scheme, which includes the conventional temporal discount scheme as special cases. The optimization problem to the new discounted objective function is proved to be solved by a value iteration algorithm; 3) We investigate the optimization problem of the *average prospect*. With one additional assumption, the solution to its optimization problem exists and a value iteration is designed to solve it; 4) For the case where the knowledge of MDP, reward and transition, is unknown, we state one algorithm to estimate the reward and transition models of underlying MDP and simultaneously learn optimal policy. For one specific prospect map (entropic map), a Q-learning like algorithm is proposed to obtain optimal policy without knowing the knowledge of MDP.

In order to avoid tedious mathematical details in general state-action spaces, we consider currently merely the MDPs with finite state-action space. However, the extensions to general space are straightforward.

This paper is organized as follows. In Sec. 2 , we briefly introduce the setups of MDPs and prospect maps, which are adapted in Sec. 3 to the MDP structure. Sec. 4 states the major theory of this paper, the discounted prospect and average prospect, whose optimal control problems are solved by value iterations under different assumptions. In Sec. 5 we discuss the existing risk-sensitive approaches and show how to represent them with specific prospect maps. Two on-line algorithms, which might be of interest for engineering-oriented audience, are stated in Sec. 6, which

is followed by experiments with simple MDPs in the final section.

2 Setup

2.1 Markov Decision Processes

A *Markov decision process* [26] is composed of a state space $\mathbf{X}$, an action space $\mathbf{A}$, a transition model Q and a reward model r. Both state and action space are assumed to be finite. The transition model $Q(y|x,a) := \mathbb{P}(X_{t+1} = y|X_t = x, A_t = a)$ denotes the probability of arriving at state y given the current state x with chosen action a at time t. We assume the transition is time-homogenous. The reward function $r(x,a) : \mathbf{X} \times \mathbf{A} \mapsto \mathbb{R}$ represents the reward obtained at state x if action a is chosen.

The policy $\pi_t(a|x)$ at time t is defined as the probability of choosing action a given state x. Let $\boldsymbol{\pi} := [\pi_0, \pi_1, \ldots]$ be the sequential policy where at time $t = 0$ the policy π_0 is used, and at $t = 1$ the policy π_1, etc. Let Π be the set of all policies. A policy is called *Markov* if for all $t \in \mathbb{N}$, π_t depends merely on x_t and is independent from all the states and actions before time t. Let Π_M denote the set of all Markov policies and Δ be the set of all one-step Markov policy. Thus $\Pi_M = \Delta^\infty$. A one-step policy $f \in \Delta$ is called *Markov deterministic*, if $f(a|x) = 1$ for some $a \in \mathbf{A}$ and $x \in \mathbf{X}$. With slight abuse of the notation, we also write f as a deterministic function $f(x) = a$. Denote the set of all one-step Markov deterministic policies by $\Delta_D \subset \Delta$. For any $\pi \in \Delta$, we define

$$r^\pi(x) := \sum_a \pi(a|x)r(x,a), P^\pi(y|x) := \sum_a \pi(a|x)Q(y|x,a) \tag{1}$$

There are usually three types of objectives functions used in the literature of MDPs, finite-stage, discounted and average reward. We summarize them as follows,

$$S_T := \sum_{t=0}^{T} r(X_t, A_t), S_\alpha := \sum_{t=0}^{\infty} \alpha^t r(X_t, A_t), \text{and } S := \lim_{T\to\infty} \frac{1}{T} S_T \tag{2}$$

where $\alpha \in [0,1)$ denotes the discount factor. Suppose we start from one given state $X_0 = x$. The optimization problem is to maximize the expected objective by selecting a policy $\boldsymbol{\pi}$:

$$\max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}^{\boldsymbol{\pi}} \left[\mathcal{S}|X_0 = x\right] \tag{3}$$

where $\mathcal{S}$ can be replaced by S_T, S_α or S[1].

2.2 Dynamic Prospect Maps

In the setup of MDPs, we apply "rewards" instead of "costs" (which are common in the literature of *Markov control processes* [16]) to model immediate outcomes and therefore in the optimization problems of MDPs (Eq. 3), objectives are to be maximized rather than minimized. To be consistent with maximizing objectives, "prospect maps" are used to name analogous nonlinear structures as *risk measures* in finance literature. Similar nomenclature can be also found in [20], where risk is replaced by *valuation*.

[1] Note that since the limit in defining the average reward S might not exist (see e.g. Example 8.1.1, [26]), the strict definition of the optimization problem of average reward should be

$$\max_{\boldsymbol{\pi}\in\Pi} \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^{\boldsymbol{\pi}} \left[S_T|X_0 = x\right].$$

Let us consider a discrete-time stochastic process $\{X_t \in \mathbf{X}\}_{t=0}^{\infty}$ and $Y_t = \{X_0, X_1, \ldots, X_t\} \in \mathbf{X}^{t+1}$. The capital letters X_t and Y_t denote random variables whereas the realizations of the random variables are denoted by normal letters, x_t and y_t, respectively. Let $\mathscr{F}_t$ denotes the set of all real-valued bounded functions on $\mathbf{X}^{t+1}$, for $t = 1, 2, \ldots$ We consider a map $R_t(v|y_t)$ such that $R(v|\cdot)$ is a real-valued bounded function on $\mathbf{X}^{t+1}$ for fixed $v \in \mathscr{F}_{t+1}$. R_t can be also viewed as a map from $\mathscr{F}_{t+1}$ to $\mathscr{F}_t$. In the following, the (in-)equalities between two functions are understood elementwise, i.e., we say $v \leq w$, if $v(x) \leq w(x)$ for all x.

In the following, we first introduce *conditional prospect maps* and then construct a *dynamic prospect measure from t to T*, $R_{t,T} : \mathscr{F}_T \mapsto \mathscr{F}_t$, by a series of conditional prospect maps $\{R_s\}_{s=t}^{T}$.

Definition 2.1. *A map $R_t : \mathscr{F}_{t+1} \mapsto \mathscr{F}_t$, $t \in \mathbb{N} \cup \{\infty\}$, is called a* conditional prospect map*, if*

I *Monotonicity. $\forall v \in \mathscr{F}_{t+1}, \forall w \in \mathscr{F}_{t+1}$, if $v \leq w$, then $R_t(v) \leq R_t(w)$.*

II *Time-consistency. For any $v \in \mathscr{F}_{t+1}$ and $\forall w \in \mathscr{F}_t$, $R_t(v + w) = w + R_t(v)$. Especially, for each $w \in \mathbb{R}$ and $v \in \mathscr{F}_{t+1}$, $R_t(v + w) = w + R_t(v)$.*

III *Centralization. $R_t(0) = 0$.*

Remarks The monotone axiom reflects the intuition that if the reward of one choice are higher than the reward of another choice, the prospect of the choice must be higher than that of the other one. The time-consistent axiom is obviously a generalization of the conditional expectation. This axiom allows the temporal decomposition (see Proposition 2.1), and together with the axiom of monotonicity make the *dynamic programming* [3] the feasible solution to the optimization problems (see Sec. 4). The axiom of centralization sets the reference point to be 0, i.e., there is no risk if there is no cost. Nevertheless, it is possible to use other reference points.

Definition 2.2. *A map $R_{t,T} : \mathscr{F}_T \mapsto \mathscr{F}_t$, $0 \leq t \leq T \in \mathbb{N} \cup \{\infty\}$, is called a* dynamic prospect map*, if there exists a series of conditional prospect maps $\{R_t\}_{s=t}^{T}$ such that*

$$R_{t,T}(v) := R_t(R_{t+1}(\ldots R_{T-1}(v) \ldots)), v \in \mathscr{F}_T.$$

Proposition 2.1. *Let $v_s \in \mathscr{F}_s$, $t \leq s \leq T$, $t, T \in \mathbb{N} \cup \{\infty\}$, $t \leq T$, and $v = \sum_{s=t}^{T} v_s \in \mathscr{F}_T$, we have*

$$R_{t,T}(v) = v_t + R_t(v_{t+1} + \ldots + R_{T-1}(v_T) \ldots)$$

Proof. Trivial using Axiom II. □

Remarks. In the literature of finance, there exist various ways to extend the CRM to a temporal structure (e.g., [9, 20, 1, 28] and references therein). The definition is usually selected based on the applications, to which the dynamic risk measures are applied. To compare their subtle differences are out of the scope of this paper. Nevertheless, there are 2 points that are remarkable: 1) in all kinds of definitions, the axiom of time-consistency is the most important component that allows the temporal decomposition as shown in Prop. 2.1; and 2) their definitions require either coherence [28] or convexity [9, 20, 1], which means that the agent has to be economically rational, i.e., risk-aversive (more discussion see Sec. 3.3). However, in some problems (especially in modeling real human behaviors), mixed risk-preference (risk-aversive at some states while risk-seeking at other states) is also a possible strategy. For instance, at gambling, some people are risk-aversive when losing money but risk-seeking when winning money. Therefore, we require neither coherence nor convexity. In this sense, our axioms are even more general than the axioms used in finance literature. Finally, in the literature of coherent risk measures, non-additive measures can be defined due to the coherency. However, in this paper we do not assume coherency in the axioms. Instead, we build the theory based on the functional spaces $\{\mathscr{F}_t\}$. Therefore, it is more accurate to use the term "map" than "measure".

3 Applying Prospect Maps in MDPs

The dynamic prospect maps introduced in Sec. 2.2 can be adapted to arbitrary temporal structures. To adapt in the structure of MDPs, we assume the prospect maps work on the state sequence $\{X_t\}_{t=0}^{\infty}$. On the other hand, since the probability of $\{X_t\}_{t=0}^{\infty}$ is controlled by policies $\boldsymbol{\pi} \in \Pi$ (together with the transition model of the MDP, Q, defined in Sec. 2.1), we assume further that the prospect maps depends on $\boldsymbol{\pi}$. Thus, the conditional prospect maps working on the MDP $(\mathbf{X}, \mathbf{A}, r, Q)$ given one policy $\boldsymbol{\pi}$ are written as $\{R_t^{\boldsymbol{\pi}}\}$.

3.1 Markov Prospect Maps for MDPs

The conditional prospect maps defined in Def. 2.1 might be dependent on the whole history, which could cause computational problems in real applications. Therefore, the prospect maps are additionally assumed to possess Markov property. Let $\mathscr{F}_B$ denote the space of all bounded functions that maps from $\mathbf{X}$ to $\mathbb{R}$.

Definition 3.1 (**Markov Prospect Map for MDPs**). *Let $\{R_t^{\boldsymbol{\pi}}\}$ be a series of conditional prospect maps defined on the MDP $(\mathbf{X}, \mathbf{A}, r, Q)$ given the policy $\boldsymbol{\pi}$. $\{R_t^{\boldsymbol{\pi}}\}$ is called* Markov*, if there exists a series of maps $\{\varrho_t : \mathscr{F}_B \mapsto \mathscr{F}_B\}$ such that*

$$R_t^{\boldsymbol{\pi}}(v(X_{t+1})|x_t, x_{t-1}, \ldots, x_0) = \varrho_t(v|x_t), \forall t \in \mathbb{N}, v \in \mathscr{F}_B$$

Remark. It is noticeable that the prospect map $\{R_t^{\boldsymbol{\pi}}\}$ depends also on Q.

From now on, we consider merely the Markov prospect maps. Thus we can write $R_t^{\boldsymbol{\pi}}$ as $R_t^{\boldsymbol{\pi}}(v(X_{t+1})|x_t)$. Furthermore, we consider merely the Markov policies $\boldsymbol{\pi} \in \Pi_M$. For a Markov random policy $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots] \in \Pi_M$, $R_t^{\boldsymbol{\pi}}(v(X_{t+1})|x_t)$ depends only on $\pi_t \in \Delta$. Hence, we can write $R_t^{\boldsymbol{\pi}}(v(X_{t+1})|x_t)$ as $R_t^{\pi_t}(v(X_{t+1})|x_t)$. For each (x_t, a_t)-pair, there exists a corresponding deterministic policy $f \in \Delta_D \subset \Delta$ satisfying $f(a_t|x_t) = 1$. Therefore, we can define for each (x_t, a_t),

$$R_t(v(X_{t+1})|x_t, a_t) := R_t^f(v(X_{t+1})|x_t) \tag{4}$$

Assumption 3.1. *We assume that the Markov prospect map $R_t^{\pi_t}$ is linear to π_t, i.e.,*

$$R_t^{\pi_t}(v(X_{t+1})|x_t) = \sum_{a \in \mathbf{A}} \pi_t(a|x_t) R_t(v(X_{t+1})|x_t, a), \forall t \in \mathbb{N}.$$

To simplify the problem, we consider merely the *time-homogeneous Markov prospect maps*, i.e., $R_t \equiv R$ for all t. Hence, $R^{\pi_t}(v(X_{t+1})|x_t)$ can be abbreviated by $R^{\pi_t}(v(X')|x)$, $x \in \mathbf{X}$, $v \in \mathscr{F}_B$ and furthermore by $R^{\pi_t}(v|x)$. Similar abbreviations are used for $R(v|x, a)$ which is a special case of $R^{\pi}(v|x)$. By Assumption 3.1, analogous to the P^{π} in Eq. 1. we obtain

$$R^{\pi}(v|x) = \sum_{a \in \mathbf{A}} \pi(a|x) R(v|x, a)$$

Then $R^{\pi}(v)$, which is defined by $R^{\pi}(v)(x) := R^{\pi}(v|x)$, is a function in the space $\mathscr{F}_B$. R^{π} can be viewed as a map from $\mathscr{F}_B$ to $\mathscr{F}_B$ itself. Since we assume the state space is finite, v can be viewed as a N-dimensional vector, where N denotes the number of states. Thus R^{π} can be understood as a map from $\mathbb{R}^N$ to $\mathbb{R}^N$ itself.

Remark. For a time-homogeneous Markov map R, Assumption 3.1 enables $R(v|x, a)$ to play the similar role as the transition model Q in MDPs. Another result of Assumption 3.1 is that for all $v \in \mathbb{R}^N$ and $x \in \mathbf{X}$, there exists a deterministic policy $f \in \Delta_D$, such that for any $\alpha \in [0, 1]$,

$$c^f(x) + \alpha R^f(v|x) = c(x, f(x)) + \alpha R(v|x, f(x)) = \min_{\pi \in \Delta} \{c^{\pi}(x) + \alpha R^{\pi}(v|x)\}.$$

3.2 Nonexpansiveness

For any time-homogeneous Markov map R, by its definition, R^π satisfies the axioms of monotonicity and time-consistency, for each $\pi \in \Delta$. Thus R^π is a *topical map* (see [12]), which satisfies, i) $R^\pi(v) \le R^\pi(w)$, whenever $v \le w \in \mathbb{R}^N$; and ii) $R^\pi(v+w) = w + R^\pi(v)$, for all $v \in \mathbb{R}^N$ and $w \in \mathbb{R}$. For each $v \in \mathscr{F}_B$, we define the *Hilbert semi-norm*[2] and *sup-norm* as follows,

$$\|v\|_H := \sup_{x,y \in \mathbf{X}} (v(x) - v(y)), \quad \|v\|_\infty := \sup_{x \in \mathbf{X}} |v(x)|.$$

Since we consider only the finite state space, v is simply an N-dimensional vector.

Suppose $F : \mathbb{R}^N \mapsto \mathbb{R}^N$ be a topical map. Then, it can be shown that F is nonexpansive under both Hilbert semi-norm and sup-norm (see Eq. 17 and 18, [12]), i.e., for all $v, w \in \mathbb{R}^N$,

$$\|F(v) - F(w)\|_H \le \|v - w\|_H, \quad \|F(v) - F(w)\|_\infty \le \|v - w\|_\infty$$

3.3 Categorization

Suppose $R^\pi(v|x)$ is a time-homogeneous Markov prospect map for some one-step Markov policy $\pi \in \Delta$. Assume furthermore $R^\pi(v|x)$ is concave with respect to v at x, i.e. for any v and $w \in \mathbb{R}^N$ and any $\beta \in [0,1]$, we have

$$R^\pi(\beta v + (1-\beta)w|x) \ge \beta R^\pi(v|x) + (1-\beta)R^\pi(w|x)$$

Note that the objective is to maximize the prospect (which will be defined Sec. 4). Suppose we have two policies π_1 and π_2 in the successive time-step which generate two outcomes v and w respectively. The concavity of $R^\pi(v|x)$ implies that the outcome of mixture of two policies, $R^\pi(\beta v + (1-\beta)w|x)$ is always preferred (due to maximization) to the mixed outcome of two single policies $\beta R^\pi(v|x) + (1-\beta)R^\pi(w|x)$. In other words, given the policy π we choose at current time step, we shall prefer mixture of two policies at the successive time-step. This shows that the corresponding risk-preference of the prospect map is *risk-aversive*. Similar result can be inferred for convex prospect maps. This categorization coincides categorization of risk-preferences judging by concavity of the utility functions in the *expected utility theory* [13]. In order to obtain a time-homogeneous risk-preference (risk-aversive or risk-seeking), the everywhere risk-preference is required. We define them as follows,

Definition 3.2. *A time-homogeneous Markov prospect map $R^\pi(v|x)$ is said to be*

(i) risk-aversive *at $x \in \mathbf{X}$, if it is concave w.r.t. v at x, and* everywhere risk-aversive*, if $R^\pi(v|x)$ is concave w.r.t. v at all $x \in \mathbf{X}$ and for all $\pi \in \Delta$.*

(ii) risk-seeking *at $x \in \mathbf{X}$, if $R^\pi(v|x)$ is convex w.r.t. v at x, and* everywhere risk-seeking*, if $R^\pi(v|x)$ is convex w.r.t. v at all $x \in \mathbf{X}$ and for all $\pi \in \Delta$.*

Remarks The categorization depends on the objective. In the CRM theory, the objective is to miminize the risk. Therefore, the categorization is opposite: concavity means risk-seeking and convexity suggests risk-aversive. Apparently under Assupmtion 3.1, if $R(v|x,a)$ is convex (concave) w.r.t. v at all (x,a)-pairs, then $R^\pi(v|x)$ is everywhere convex (everywhere concave). Several existing risk maps (see Sec. 5) in the literature confirm also the above defined categorizations.

One widely used family of prospect maps, the *coherent prospect maps*, is worth mentioning.

Definition 3.3. *A time-homogeneous Markov prospect map $R^\pi(v|x)$ is said to be* coherent *if for all $\lambda > 0$, $R^\pi(\lambda v|x) = \lambda R^\pi(\lambda v|x)$ for all $\pi \in \Delta$, $v \in \mathbb{R}^N$ and $x \in \mathbf{X}$.*

[2]Here we follow the terminology in [12, 24], whereas the same semi-norm is called *span semi-norm* in [26, 15].

4 Discounted and Average Prospect

4.1 Finite-stage Prospect

According to the definition of dynamic prospect maps (Def. 2.2), we define the T-stage total prospect as follows,

$$J_T(x, \boldsymbol{\pi}) := R^{\boldsymbol{\pi}}_{0,T}(\sum_{t=0}^{T} r(X_t, A_t)|X_0 = x) \tag{5}$$

Suppose the prospect map $\{R_t\}$ under consideration is time-homogeneous and Markov. By Prop. 2.1, we have the following decomposition

$$J_T(x, \boldsymbol{\pi}) = r^{\pi_0}(x) + R^{\pi_0}_{X_0=x}\left[r^{\pi_1}(X_1) + R^{\pi_1}_{X_1}\left[r^{\pi_2}(X_2) + \ldots + R^{\pi_{T-1}}_{X_{T-1}}[r^{\pi_T}(X_T)]\ldots\right]\right]$$

where the short notation $R^{\pi_t}_{X_t}(v(X_{t+1})) := R^{\pi_t}(v(X_{t+1})|X_t)$ is used. The optimization problem of this objective function is to maximize the T-stage total prospect among all Markov random policies, i.e.,

$$J^*_T(x) = \max_{\boldsymbol{\pi}\in\Pi_M} J_T(x, \boldsymbol{\pi})$$

Suppose Assumption 3.1 holds true. Obviously, the optimization problem can be solved by *dynamic programming*, i.e., we start from

$$V_T(x) = \max_{\pi\in\Delta} r^{\pi}(x) = \max_{a\in\mathbf{A}} r(x, a)$$

Then we calculate backwards, for $t = T-1, T-2, \ldots, 0$,

$$V_t(x) = \max_{\pi\in\Delta}\{r^{\pi}(x) + R^{\pi}(V_{t+1}|x)\} = \max_{a\in\mathbf{A}}\{r(x, a) + R(V_{t+1}|x, a)\}$$

It is easy to verify that $V_0(x) = J^*_T(x)$.

4.2 Discounted Total Prospect

Let $\alpha \in [0, 1)$ denote the discount factor. Suppose Assumption 3.1 holds true. We use the discounted T-stage prospect as follows,

$$J_{\alpha,T}(x, \boldsymbol{\pi}) := r^{\pi_0}(x) + \alpha R^{\pi_0}_{X_0=x}\left[r^{\pi_1}(X_1) + \alpha R^{\pi_1}_{X_1}\left[r^{\pi_2}(X_2) + \ldots + \alpha R^{\pi_{T-1}}_{X_{T-1}}[r^{\pi_T}(X_T)]\ldots\right]\right] \tag{6}$$

and the discounted total prospect as

$$J_{\alpha}(x, \boldsymbol{\pi}) := \lim_{T\to\infty} J_{\alpha,T}(x, \boldsymbol{\pi}) \tag{7}$$

Thus, the optimization problem for discounted total prospect is

$$J^*_{\alpha}(x) := \sup_{\boldsymbol{\pi}\in\Pi_M} J_{\alpha}(x, \boldsymbol{\pi})$$

We first prove that the limit exists in Eq. 7. Given $\pi \in \Delta$, we define the map $F^{\pi}_{\alpha} : \mathbb{R}^N \mapsto \mathbb{R}^N$ as $F^{\pi}_{\alpha}(v|x) := r^{\pi}(x) + R^{\pi}(v|x)$, $v \in \mathbb{R}^N$ and $x \in \mathbf{X}$. For any $\boldsymbol{\pi} \in \Pi_M$, define

$$F^{\boldsymbol{\pi}}_{\alpha,T}(v) := F^{\pi_0}_{\alpha}(F^{\pi_1}_{\alpha}(\ldots F^{\pi_{T-1}}_{\alpha}(v)\ldots)).$$

Proposition 4.1. *For any $\boldsymbol{\pi} \in \Pi_M$, i) the limit in Eq. 7 exists; ii)* $\lim_{T\to\infty} F^{\boldsymbol{\pi}}_{\alpha,T}(v) = J_\alpha(\boldsymbol{\pi})$, $\forall v \in \mathbb{R}^N$.

Proof. (i) Since r is bounded for finite state and action spaces, there exists a number $M < \infty$ such that $|r(x,a)| \le M$ for all (x,a). Hence, by monotonicity and additive property of R,

$$-\alpha^{T+1} M \le J_{\alpha,T+1}(x,\boldsymbol{\pi}) - J_{\alpha,T}(x,\boldsymbol{\pi}) \le \alpha^{T+1} M$$

which implies $|J_{\alpha,T+1}(x,\boldsymbol{\pi}) - J_{\alpha,T}(x,\boldsymbol{\pi})| \to 0$ as $T \to \infty$.

(ii) Since $v \in \mathbb{R}^N$, $v - r^\pi$ is also bounded for all $\pi \in \Delta$. Let M' be the upper bound such that $\|v - r^\pi\|_\infty \le M'$. Hence,

$$r^\pi - M' \le v \le r^\pi + M' \Rightarrow -M'\alpha^T \le F^{\boldsymbol{\pi}}_{\alpha,T}(v) - J_{\alpha,T}(x,\boldsymbol{\pi}) \le M'\alpha^T$$

Using the conclusion of (i), we have $\lim_{T\to\infty} F^{\boldsymbol{\pi}}_{\alpha,T}(v) = J_\alpha(\boldsymbol{\pi})$, $\forall v \in \mathbb{R}^N$. □

Discussion The trivial extension of the classical discounted MDP (cf. S_α in Eq. 2) is as follows,

$$D_\alpha(x,\boldsymbol{\pi}) := R^{\boldsymbol{\pi}}_{0,\infty}(\sum_{t=0}^{\infty} \alpha^t r(X_t, A_t)|X_0 = x)$$

Using the time-consistency property of prospect maps, we have the following decomposition

$$D_\alpha(x,\boldsymbol{\pi}) = r^{\pi_0}(x) + R^{\pi_0}_{X_0=x}\left[\alpha r^{\pi_1}(X_1) + R^{\pi_1}_{X_1}\left[\alpha^2 r^{\pi_2}(X_2) + \ldots + R^{\pi_{T-1}}_{X_{T-1}}[\alpha^T r^{\pi_T}(X_T) + \ldots] \ldots\right]\right]$$

We have the following observations:

- We can prove analogously as in Prop. 4.1(i) that D_α is well-defined.
- If the prospect map R is coherent, then D_α is equivalent to J_α (cf. Eq. 7), the discounted total prospect under our definition. Therefore, D_α defined for any coherent prospect map is merely special cases of our definition. Especially, the discounted total reward in classical MDPs is a special case of the discounted total prospect, since it is coherent.
- For general prospect maps, there might not exist a stationary policy that D_α, as proved by Chung & Sobel (1987) [7] for entropic prospect maps, which are not coherent. We can prove analogous statements as Theorem 4 in [7] for arbitrary non-coherent prospect maps.
- Ruszczyński (2010) [28] uses D_α as the objective function, which was solved by a value iteration algorithm. However, in the proof of the value iteration algorithm, he uses the representation theorem which is valid merely for coherent prospect maps. On the contrary, we will see later that the objective J_α allows a value iteration algorithm for arbitrary prospect maps.

Contracting Map Given a function $u \in \mathbb{R}^N$ and $x \in \mathbf{X}$, consider the following map

$$F_\alpha(u|x) := \max_{\pi\in\Delta} F^\pi_\alpha(u|x) = \max_{a\in\mathbf{A}} [r(x,a) + \alpha R(u|x,a)] \text{ (under Assumption 3.1)}$$

Now we prove the key property: F_α is a contracting map.

Lemma 4.1. *Suppose Assumption 3.1 holds true. Then F_α is a contracting map under sup-norm, i.e., $\|F_\alpha(u) - F_\alpha(v)\|_\infty \le \alpha\|u - v\|_\infty$, for all u and $v \in \mathbb{R}^N$.*

Proof. Under Assumption 3.1, there exist deterministic policy f and g satisfying

$$F_\alpha(u|x) = r(x, f(x)) + \alpha R(u|x, f(x)), \quad F_\alpha(v|x) = r(x, g(x)) + \alpha R(u|x, g(x))$$

By definition, we have for all $x \in \mathbf{X}$,

$$\begin{aligned} F_\alpha(u|x) - F_\alpha(v|x) \leq & r(x, f(x)) + \alpha R(u|x, f(x)) - r(x, f(x)) - \alpha R(v|x, f(x)) \\ = & \alpha \left[R(u|x, f(x)) - R(v|x, f(x)) \right] \leq \alpha \|u - v\|_\infty \end{aligned}$$

where the last inequality is due to the nonexpansiveness of R. Exchanging v and w, we have

$$F_\alpha(v|x) - F_\alpha(u|x) \leq \alpha \|v - u\|_\infty$$

Thus, $|F_\alpha(u|x) - F_\alpha(v|x)| \leq \alpha\|u - v\|_\infty$ for all $x \in \mathbf{X}$, which implies $\|F_\alpha(u) - F_\alpha(v)\|_\infty \leq \alpha\|u - v\|_\infty$ □

Value iteration We state the following algorithm:

1. select one $v_0 \in \mathbb{R}^N$, $t = 0$;
2. calculate $v_{t+1} = F_\alpha(v_t)$, $f_t = \arg\max F_\alpha(v_t)$
3. if $\|v_{t+1} - v_t\|_\infty < \epsilon$, stop; otherwise, $t \Leftarrow t + 1$ and goto step 2.

Since F_α is a contracting map, due to the Banach contraction mapping principle, we conclude that for all $v \in \mathbb{R}^N$, $v_t \to v^*$ and $f_t \to f^*$, as $t \to \infty$, where v^* is the fixed point of F_α such that $F_\alpha(v^*) = v^*$ and f^* denotes the corresponding policy. The final step is to prove $v^* = J_\alpha^*$ with the following theorem.

Theorem 4.1. *Suppose Assumption 3.1 holds true. For any $v \in \mathbb{R}^N$, i) if $v \geq F_\alpha(v)$, then $v \geq J_\alpha^*$; ii) If $v \leq F_\alpha(v)$, then $v \leq J_\alpha^*$; iii) if $v = F_\alpha(v)$, then $v = J_\alpha^*$.*

Proof. (i) Consider a Markov policy $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots]$. $v \geq F_\alpha(v)$ implies that for any $\pi \in \Delta$,

$$v \geq F_\alpha(v) \geq r^\pi + \alpha R^\pi(v)$$

We apply above inequality recursively,

$$v \geq r^{\pi_0} + \alpha R^{\pi_0}(v) \geq r^{\pi_0} + \alpha R^{\pi_0}(r^{\pi_1} + \alpha R^{\pi_1}(v)) \geq \ldots \geq J_\alpha(\boldsymbol{\pi})$$

Since $\boldsymbol{\pi}$ is arbitrary, above inequality implies $v \geq \sup J_\alpha(\boldsymbol{\pi}) = J_\alpha^*$.

(ii) Under Assumption 3.1, there exists an $f \in \Delta_D$ such that $F_\alpha(v|x) = r(x, f(x)) + \alpha R(v|x, f(x))$. Write $r^f(x) := r(x, f(x))$ and $R^f(v|x) := R(v|x, f(x))$. Since $v \leq F_\alpha(v)$, we have

$$v \leq F_\alpha^f(v) = r^f + \alpha R^f(v) \leq r^f + \alpha R^f(r^f + \alpha R^f(v)) \leq \ldots \leq J_\alpha(f^\infty) \leq J_\alpha^*$$

where we apply the monotonicity of R^f recursively. Due to Prop. 4.1(ii), for any $\boldsymbol{\pi} \in \Pi_M$, $J_\alpha(\cdot, \boldsymbol{\pi})$ exists. (i) + (ii) implies (iii). □

4.3 Average Prospect

Analogous to the average reward S defined in Eq. 2, we consider the following average prospect,

$$J(x, \boldsymbol{\pi}) := \liminf_{T\to\infty} \frac{1}{T} J_T(x, \boldsymbol{\pi}), \boldsymbol{\pi} \in \Pi_M, x \in \mathbf{X}$$

where $J_T(x, \boldsymbol{\pi})$ is defined in Eq. 5. Here "lim inf" is used to avoid the case where the limit of $\frac{1}{T} J_T$ does not exist (see e.g., Example 8.1.1, [26]). The optimization problem of average prospect is therefore,

$$J^*(x) = \sup_{\boldsymbol{\pi}\in\Pi_M} J(x, \boldsymbol{\pi})$$

Suppose there is a pair $(h, \rho), h \in \mathbb{R}^N, \rho \in \mathbb{R}$, which satisfies the following equation

$$\rho + h(x) = \max_{\pi\in\Delta} [r^\pi(x) + R^\pi(h|x)] \tag{8}$$

This equation is called *average prospect optimality equation* (APOE). Under Assumption 3.1, there exists a deterministic function $f \in \Delta$ such that

$$\rho + h(x) = \max_{a\in\mathbf{A}} [r(x, a) + R(h|x, a)] = r(x, f(x)) + R(h|x, f(x))$$

Define operator F^π as

$$F^\pi(v) := r^\pi + R^\pi(v), v \in \mathbb{R}^N$$

Let $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots] \in \Pi_M$ be an arbitrary random Markov policy. Define

$$F_T^{\boldsymbol{\pi}}(v) := F^{\pi_0}(F^{\pi_1}(\ldots F^{\pi_{T-1}}(v)\ldots)) \tag{9}$$

Lemma 4.2. *Suppose the Assumption 3.1 holds true and the APOE has a solution (h, ρ), $h \in \mathbb{R}^N$ and $\rho \in \mathbb{R}$. Let f be the deterministic policy found in the APOE. Then $\rho = J(x, f^\infty) = J^*(x)$, for all $x \in \mathbf{X}$.*

Proof. We prove $\rho = J(x, f^\infty)$ first. Define an operator $F : \mathbb{R}^N \mapsto \mathbb{R}^N$ as follows,

$$F(v) := r^f + R^f(v), v \in \mathbb{R}^N$$

and $F^{t+1}(v) := F(F^t(v))$, $t = 1, 2, \ldots$. Hence, due to the nonexpansiveness of R^f, we have

$$\|J_T(f^\infty) - F^T(h)\|_\infty \le \|r^f - h\|_\infty \Rightarrow \lim_{T\to\infty} \left(\frac{1}{T} J_T(f^\infty) - \frac{1}{T} F^T(h)\right) = 0 \tag{10}$$

On the other hand, by APOE, we have

$$F^T(h) = F^{T-1}(r^f + R^f(h)) = F^{T-1}(h) + \rho = \ldots = h + T \cdot \rho$$

Hence, $\lim_{T\to\infty} \frac{1}{T} F^T(h) = \rho$. Together with Eq. 10, we obtain $\lim_{T\to\infty} \frac{1}{T} J_T(f^\infty) = \rho$.

Now we prove that $J(x, \boldsymbol{\pi}) \le \rho$ for any $\boldsymbol{\pi} \in \Pi_M$ and all $x \in \mathbf{X}$. By AROE, we have for all $\pi \in \Delta$,

$$\rho + h \ge r^\pi + R^\pi(h) \tag{11}$$

Let $\boldsymbol{\pi} \in \Pi_M$ be any Markov random policy. Then $F_T^{\boldsymbol{\pi}}$ defined in Eq. 9 satisfies,

$$\|J_T(\boldsymbol{\pi}) - F_T^{\boldsymbol{\pi}}(v)\|_\infty \leq \|r^f - v\|_\infty \Longrightarrow \lim_{T\to\infty}\left(\frac{1}{T}J_T(\boldsymbol{\pi}) - \frac{1}{T}F_T^{\boldsymbol{\pi}}(v)\right) = 0 \tag{12}$$

By Eq. 11, we have

$$\begin{aligned}
F_T^{\boldsymbol{\pi}}(h) =& F_{T-1}^{\boldsymbol{\pi}}(r^{\pi_{T-1}} + R^{\pi_{T-1}}(h)) \\
\leq& F_{T-1}^{\boldsymbol{\pi}}(\rho + h) = F_{T-1}^{\boldsymbol{\pi}}(h) + \rho \\
\leq& \ldots \leq h + T\cdot\rho
\end{aligned}$$

which implies

$$\liminf_{T\to\infty}\frac{1}{T}F_T^{\boldsymbol{\pi}}(h) \leq \rho \stackrel{Eq.\ 12}{\Longrightarrow} \liminf_{T\to\infty}\frac{1}{T}J_T(\boldsymbol{\pi}) \leq \rho$$

□

Now the question is to find proper assumptions that can guarantee the existence of the solutions of the APOE. Assumption 3.1 is not sufficient to take this burden. Recall that $\|\cdot\|_H$ denotes Hilbert semi-norm defined in Sec.3.2. We further assumes

Assumption 4.1. *There exists an integer K and a real number $\beta \in [0,1)$ such that for all deterministic policy $\boldsymbol{\pi} = [f_0, f_1, \ldots, f_{K-1}] \in \Delta_D^K$*

$$\|R^{\boldsymbol{\pi}}(u) - R^{\boldsymbol{\pi}}(v)\|_H \leq \beta\|u - v\|_H, \forall u, v \in \mathbb{R}^N$$

where $R^{\boldsymbol{\pi}}(\cdot) := R^{f_0}(R^{f_1}\ldots R^{f_{K-1}}(\cdot)\ldots)$.

Define the operator, $F : \mathbb{R}^N \mapsto \mathbb{R}^N$,

$$F(v|x) := \max_{a\in\mathbf{A}}\left\{r(x,a) + R(v|x,a)\right\}, \quad F^t(v) := F(F^{t-1}(v)), t = 1, 2, \ldots \tag{13}$$

Proposition 4.2. *If Assumption 3.1 and 4.1 hold true, then $\|F^K(u) - F^K(v)\|_H \leq \beta\|u - v\|_H$, for all $u, v \in \mathbb{R}^N$.*

Proof. Let $F_K^{\boldsymbol{\pi}}$ be as defined in Eq. 9. There must be two policies $\boldsymbol{\pi}_u = [f_0, f_1, \ldots, f_{K-1}], \boldsymbol{\pi}_v = [g_0, g_1, \ldots, g_{K-1}] \in \Delta_D^K$ satisfying $F_K^{\boldsymbol{\pi}_u}(u) = F^K(u)$ and $F_K^{\boldsymbol{\pi}_v}(v) = F^K(v)$ respectively.

$$\begin{aligned}
F^K(u) - F^K(v) \leq& F_K^{\boldsymbol{\pi}_u}(u) - F_K^{\boldsymbol{\pi}_u}(v) \\
=& R^{f_0}(c^{f_1} + R^{f_1}(\ldots + R^{f_{K-1}}(u)\ldots)) - R^{f_0}(c^{f_1} + R^{f_1}(\ldots + R^{f_{K-1}}(v)\ldots)) \\
\text{(Prop. 2.1)} =& R^{f_0}(R^{f_1}(\ldots R^{f_{K-1}}(\sum_{t=1}^{K-1}c^{f_t} + u)\ldots)) - R^{f_0}(R^{f_1}(\ldots R^{f_{K-1}}(\sum_{t=1}^{K-1}c^{f_t} + v)\ldots))
\end{aligned}$$

Exchange u and v, we have $F^K(v) - F^K(u) \leq F_K^{\boldsymbol{\pi}_v}(v) - F_K^{\boldsymbol{\pi}_v}(u)$. Thus,

$$\begin{aligned}
\|F^K(u) - F^K(v)\|_H \leq& \max_{\boldsymbol{\pi}\in\Delta_D^K}\|F_K^{\boldsymbol{\pi}}(u) - F_K^{\boldsymbol{\pi}}(v)\| \\
=& \max_{\boldsymbol{\pi}\in\Delta_D^K}\|R^{\boldsymbol{\pi}}(\sum_{t=1}^{K-1}c^{\pi_t} + u) - R^{\boldsymbol{\pi}}(\sum_{t=1}^{K-1}c^{\pi_t} + v)\| \leq \beta\|u - v\|_H
\end{aligned}$$

□

Theorem 4.2. *Suppose Assumption 3.1 and 4.1 hold true. Then there exist $\rho \in \mathbb{R}$ and $h \in \mathbb{R}^N$ satisfying the APOE (Eq. 8) and furthermore $\lim_{t\to\infty} F^t(v_0) = h$ for all $v_0 \in \mathbb{R}^N$, where F^t is defined in Eq. 13.*

Proof. Let K be the integer in Assumption 4.1. Then by Prop. 4.2, we have

$$\|F^K(u) - F^K(v)\|_H \le \beta\|u - v\|_H, \forall u, v \in \mathbb{R}^N$$

Starting from an arbitrary point $v_0 \in \mathbb{R}^N$, the iteratively computed sequence $v_{t+1} = Fv_t$, $t = 0, 1, 2, \ldots$, satisfies

$$\|v_{tK+1} - v_{tK}\|_H \le \beta\|v_{(t-1)K+1} - v_{(t-1)K}\|_H \le \ldots \le \beta^t\|v_1 - v_0\|_H$$

Thus, $\|v_{tK+1} - v_{tK}\|_H \to 0$ as $t \to \infty$. Since F is nonexpansive in Hilbert semi-norm, we have $\|v_{t+1} - v_t\|_H \le \|v_{[t/K]\cdot K+1} - v_{[t/K]\cdot K}\|_H$, where $[t/K] := \max\{i \in \mathbb{N}, iK \le t\}$. Thus, $\forall t \in \mathbb{N}$,

$$\|v_{t+1} - v_0\|_H \le \sum_{i=0}^{t}\|v_{i+1} - v_i\|_H \le \sum_{i=0}^{t} K\beta^{[i/K]}\|v_1 - v_0\|_H \le \frac{K}{1-\beta}\|v_1 - v_0\|_H$$

which implies that for all t, v_t is bounded and therefore in the space $\mathbb{R}^N$, i.e.,

$$h := \lim_{t\to\infty} v_t \in \mathbb{R}^N \text{ exists}$$

Obviously, since h is bounded, $F(h)$ is bounded, as well as $F(h) - h$. Hence, due to the fact $\|Fh - h\|_H = 0$, there exists a finite $\rho \in \mathbb{R}$ satisfying

$$\rho + h(x) = F(h|x) = \min_{a\in\mathbf{A}(x)} \{c(x, a) + R(h|x, a)\}$$

which is the AROE.

□

Remarks In classical MDPs with finite state-action space, $R(v|x, a) = \sum_{y\in\mathbf{X}} Q(y|x, a)v(y)$. It can be shown (Theorem 8.5.3, [26]) that if for each $a \in \mathbf{A}$, $Q(y|x, a)$ is recurrent (irreducible and aperiodic), then there exists some integer K and one state $y \in \mathbf{X}$ such that for all K-stage deterministic policies $\boldsymbol{\pi} \in \Delta_D^K$, $P^{\boldsymbol{\pi}}(y|x) > 0$, for all $x \in \mathbf{X}$, and furthermore Assumption 4.1 holds true (Theorem 8.5.2 and 6.6.2, [26]). Therefore, for classical MDPs, Assumption 4.1 is equivalent to the conventional assumption of recurrence.

However, for general prospect maps, how to easily verify whether Assumption 4.1 is satisfied is still an open question. Some insights were given in [12], where the properties of general topical maps are investigated via the associated graph, $\mathcal{G}(R)$, of the topical map R. They found some sufficient conditions (strongly connectedness of the associated graph) for guaranteeing the existence of fixed point in Hilbert space. However, we find that the conditions would fail for entropic maps (for definition see Eq. 15) when $\lambda < 0$. We should leave this job as future work.

Value iteration Based on Theorem 4.2, we state the following algorithm:

1. select one $v_0 \in \mathbb{R}^N$, $t = 0$;
2. calculate $v_{t+1} = F(v_t)$; $f_t = \arg\max F(v_t)$
3. if $\|v_{t+1} - v_t\|_H < \epsilon$, stop; otherwise $t \Leftarrow t + 1$ and goto step 2.

Theorem 4.2 guarantees that $v_t \to h$, $v_{t+1} - v_t \to \rho$ and $f_t \to f^*$ as $t \to \infty$, where h and ρ are the solutions to the APOE and f^* denotes the optimal policy of the average prospect problem. More specifically, ρ is equivalent to J^*.

5 Examples

There exist several important risk-sensitive maps in the literature of economics, finance and control theory. Most of them can be adapted to the framework we introduced above. We assume the prospect maps under consideration are time-homogeneous and Markov. Suppose Assumption 3.1 holds true. Therefore the form of $R(v|x, a)$ determines the prospect map that we select. In the following part, we simply state the form of conditional prospect map for each specific prospect map. There are only a few exceptions in literature where the maps they used are not within our framework, e.g. value at risk [17] and mean-variance trade-off for entire MDPs [18]. In fact, in those exceptions, dynamic programming can not be applied due to lack of time-consistency and therefore, exact solutions to the optimization problems are usually computationally infeasible for high dimensional state-action space.

Classical MDPs [4, 26]

$$R(v|x, a) := \mathbb{E}^{Q}_{x,a}\left[v(X')\right] = \sum_{y \in \mathbf{X}} Q(y|x, a)v(y) \tag{14}$$

It is easy to verify that R is coherent and linear to v (therefore risk-neutral).

Entropic maps The name is taken from the literature of CRM [11]. It is also lengthily researched in the operations research (e.g., Borkar (2002) [5] and references therein) and the control theory (e.g., Coraluppi & Marcus (2000), [8] and references therein), due to its good properties.

$$R(v|x, a) := \frac{1}{\lambda} \log \mathbb{E}^{Q}_{x,a}\left[\exp(\lambda v(X'))\right] = \frac{1}{\lambda} \log \left\{ \sum_{y \in \mathbf{X}} Q(y|x, a) \exp(\lambda v(y)) \right\} \tag{15}$$

where the risk-sensitive parameter $\lambda \in \mathbb{R}$ controls the risk-preference of the risk map R: if $\lambda > 0$, R is everywhere convex and therefore everywhere risk-seeking; if $\lambda < 0$, R is everywhere concave and therefore everywhere risk-aversive. It can be also shown that

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \log \left\{ \sum_{y} Q(y|x, a) \exp(\lambda v(y)) \right\} = \sum_{y} Q(y|x, a)v(y)$$

which is exactly the conditional map of classical MDPs. Besides, it has connection to the mean-variance trade-off scheme via the following Taylor expansion at $\lambda = 0$,

$$\frac{1}{\lambda} \log \mathbb{E} \exp(\lambda Z) = \mathbb{E}Z + \lambda \mathrm{Var}(Z) + O(\lambda)$$

where Z denotes arbitrary random variable. Suppose that risk is measured by variance. The objective is to maximize R^{π} (see Sec. 4.1). Therefore, if $\lambda < 0$, the variance is avoided, the agent is intuitively risk-aversive. Conversely, if $\lambda > 0$, the variance is preferred, the agent is intuitively risk-seeking. These intuitions coincide the categorization based on the convexity (concavity) of R.

Remark. There are some literature (e.g., Borkar (2002) [5]) that do not satisfy the Assumption 3.1 we make. Instead of $R^{\pi}(v|x) = \sum_{a} \pi(a|x)R(v|x, a)$, they define

$$R^{\pi}(v|x) := \frac{1}{\lambda} \log \left\{ \sum_{y \in \mathbf{X}} P^{\pi}(y|x) \exp(\lambda v(y)) \right\}$$

i.e., π is inside the log function rather than outside the log function as in our definition. However, the optimal policy they find is still deterministic and is equivalent to the optimal policy according to our definition. In this sense, there is no essential difference between their definition and ours.

Robust maps Iyengar (2005) [19] invented the framework of *robust dynamic programming*. He argues that in some applications the transition model Q can not be inferred exactly. Instead, he employs a set of transition probabilities, $\mathcal{P}$, which contains all possible "ambiguous" transition probabilities. In order to gain the "robustness", the worst cost (i.e., lowest reward) is considered, adapted in our framework, i.e.,

$$R(v|x,a) := \inf_{Q\in\mathcal{P}} \mathbb{E}^{Q}_{x,a} v(X') = \inf_{Q\in\mathcal{P}} \sum_{y\in\mathbf{X}} Q(y|x,a)v(y)$$

It is apparent that R is coherent. We can verify that R is everywhere concave and therefore risk-aversive, which coincides the intuition that the worst scenario is considered. One special case of the robust dynamic programming was the *minimax control* (details see Coraluppi, 1997 [31]), which also considers the worst scenario and can be used only in finite state space,

$$R(v|x,a) := \min_{Q(y|x,a)>0} v(y)$$

Conditional average value at risk has important applications in finance (see e.g. [27]). Adapted to our framework, it can be defined as,

$$R(v|x,a) = \sup_{u\in\mathbb{R}} \left\{ u + \frac{1}{\alpha}\mathbb{E}^{Q}_{x,a}\left[(v(X') - u)_{+}\right] \right\} \tag{16}$$

where $(z)_+$ denotes $z \vee 0$. R is coherent and everywhere concave. Therefore, this prospect map is risk-aversive.

Mean-semideviation maps [25] This map considers only the trade-off between the one-step conditional mean and semideviation (see Eq. below) rather than the deviation of the whole Markov chain (see [30, 10]).

$$R(v|x,a) := \mathbb{E}^{Q}_{x,a}[v(X')] + \lambda \mathbb{E}^{Q}_{x,a}\left[(v(X') - \mathbb{E}^{Q}_{x,a}[v(X')])^r\right]^{1/r} \tag{17}$$

where $r \geq 1$ and $\lambda \in \mathbb{R}$ denotes the risk-preference parameter which controls the risk-preference of R: if $\lambda < 0$, R is risk-aversive; if $\lambda > 0$, R is risk-seeking. It can be shown that R is concave with negative λ whereas convex with positive λ. This map was used by Gosavi (2006) [14] (with setting $r + 2$) to approximate the mean-variance trade-off scheme defined in [10].

Probability weighting maps Consider the following map

$$R(r|x,a) := \mathbb{E}^{Q^w}_{x,a}[u(r(X'))] := \sum_{y\in\mathbf{X}} w(Q(y|x,a))u(r(y))$$

where $u(\cdot)$ and $w(\cdot)$ denote utility function and probability weighting function respectively. u is assumed to be a monotonically increasing function and satisfying $u(0) = 0$. $w(\cdot)$ is also monotonically increasing and satisfies $w(0) = 0$ and $w(1) = 1$. Note that for general utility functions, the map R constructed above satisfies the axiom of monotonicity and centralization, but not necessarily the axiom of time-consistency. In order to amend the problem, we replace immediate rewards $r(x,a)$ in defining total prospect (see Eq. 5 and 6) with its utility $u(r(x,a))$.

This map has a long history in economics to model mixed risk-preference, which is determined by the setting of utility function and probability weighting functions (a nice review see [32]). However, in economics, it is only used to model single decision problem. Here we generalize it to adapt to the temporal structure of MDPs.

Choquet integral was applied in Chateauneuf & Cohen (2008) [6] to model the *subjective expected utility* [29]. The Choquet integral is determined by a non-additive measure μ, which satisfies the monotonicity and centralization. Furthermore, Choquet integral is coherent. Chateauneuf & Cohen (2008) focus on the one-step decision problem. However, it is trivial to extend the theory to the MDP structure. In fact, Choquet integral can be viewed as a coherent prospect map, which is a special case within our general framework.

6 Reinforcement Learning Approaches

In real applications, the transition model Q and reward function r are not known before exploring the system and collecting samples. Reinforcement learning (RL) approaches like *temporal-difference learning* [33] and *Q-learning*, are popular online algorithms where Q and r are not required. These approaches belong to a the class of *stochastic approximation algorithms* [22], whose key point is that the conditional prospect map $R(v)$ is linear in the transition model Q. The conditional prospect map of classical MDPs has this property (cf. Eq. 14). However, the transition model Q is usually necessary to compute the conditional prospect map and therefore the stochastic approaches like Q-learning are in general not applicable except for some specific maps that can be transformed to an equivalent map which is linear in Q. This is, for example, the case for entropic maps.

Q-Learning for Entropic Measure We consider only the risk-aversive case, i.e. $\lambda < 0$, while the risk-seeking case can be dealt with similarly. Substitute Eq. 15 (with discount factor α) into the value iteration,

$$v_{t+1}(x) = \max_a \left\{ r(x,a) + \frac{\alpha}{\lambda} \log \mathbb{E}^Q_{x,a}[e^{\lambda v_t}] \right\} \Leftrightarrow e^{\frac{\lambda}{\alpha} v_{t+1}(x)} = \min_a \left\{ e^{\frac{\lambda}{\alpha} r(x,a)} \mathbb{E}^Q_{x,a}[e^{\lambda v_t}] \right\}$$

Let $w := \exp(\frac{\lambda}{\alpha} v)$. The above equation is equivalent to $w_{t+1}(x) = \min_a \left\{ e^{\frac{\lambda}{\alpha} r(x,a)} \mathbb{E}[(w_t)^\alpha | x, a] \right\}$, which is linear in the transition model Q. Observed the state x_t, action a_t, the reward r_t at time t and the successive state x_{t+1}, the update rule of Q-learning for w (instead of v) is

$$q_{t+1}(x_t, a_t) = q_t(x_t, a_t) + \beta_t \left[e^{\frac{\lambda}{\alpha} r_t} \min_a \left(q_t(x_{t+1}, a) \right)^\alpha - q_t(x_t, a_t) \right] \tag{18}$$

where β_t denotes the learning rate.

Model-based Approaches for General Prospect Maps We introduce an algorithm similar to the dyna-Q approach [33]. Repeat the following procedure until convergence:

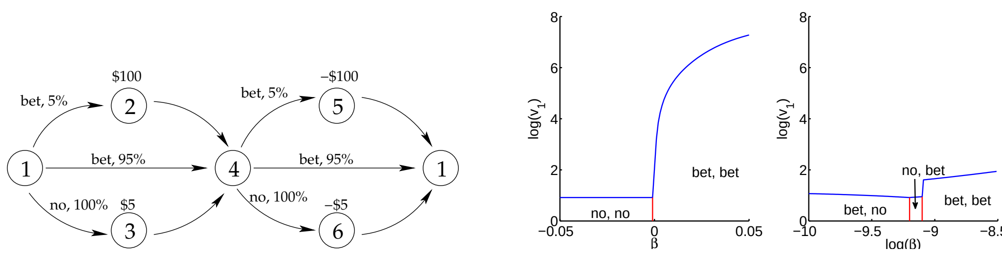
1. Given data (x_t, a_t, x_{t+1}, r_t) update the model estimates $\hat{Q}^{(t)}$ and $\hat{r}^{(t)}$
2. Update the Q-value at (x_t, a_t) based on the estimated models $\hat{Q}^{(t)}$ and $\hat{r}^{(t)}$ by $q(x_t, a_t) = \hat{r}^{(t)}(x_t, a_t) + \rho(\gamma \max_a q(x_{t+1}, a) | x_t, a_t, \hat{Q}^{(t)})$
3. Perform k additional updates: choose k state-action pairs at random and update them according to the same rule: $q(x_k, a_k) = \hat{r}^{(t)}(x_k, a_k) + \rho(\max_a \gamma q(x_{k+1}, a) | x_k, a_k, \hat{Q}^{(t)})$
4. Choose an action a_{t+1} at state x_{t+1}, based on the softmax or ϵ-greedy policy. Go to Step 1.

7 Experiments with Simple MDPs

By presenting two experiments, we aim to illustrate by the first experiment the capability of modeling mixed risk-preference via designing new prospect map under our framework, and by the second experiment to verify the effectiveness of the Q-learning algorithm introduced in Sec. 6.

Experiment 1: Sequential Betting Game The prospect simply expresses preferences of the agent and there is no right or wrong choice. The framework we introduced here is a set of options for the agent and we want to show in this first experiment that this set contains useful options. Furthermore, we compare the solutions of optimizing different maps to see whether there is a significant difference of resulting optimal policies.

The task is a game that has two stages and at both stages the agent faces two options: he can either bet or do nothing (denoted with *bet* and *no* in Fig. 1(a)). At the 1st stage (State 1 in Fig. 1(a)), if "bet" is chosen then the agent will be rewarded with \$100 with a 5% chance and \$0 with 95%; if "no" is chosen then he gains \$5. At the 2nd stage (State 4 in Fig. 1(a)), if "bet" is chosen then the agent will suffer a loss of \$100 with 5% chance and \$0 with 95%; if "no" is chosen, then he loses \$5. From stage 2 he goes back to stage 1 to repeat the game. The discounting factor is set to 0.99.



(a) MDP

(b) Value of State 1 vs λ (in figure is β).

Figure 1: Sequential Betting Game

The betting game expresses in a way the essence of risk choices: risk-neutral behavior will be indifferent between the options "bet" and "not bet" as the mean reward for the two options is the same in both stages. It is thus a question of risk-preference what should be done. A risk-seeking person or company will prefer the "bet" option. A risk-aversive user will prefer "not bet". Also a mixed risk-preference might be preferable. For example, people might be happy to bet when they can't lose money (stage 1) but might not be happy to bet if they have to pay if they lose (stage 2).

The total reward expresses the risk-neutral option and a user that is indifferent to risk should chose this criterion. The entropic map can both encode risk-aversion and risk-seeking depending on its parameter λ. We calculated the optimal policy using entropic maps for different values of λ with the value iteration (cf. Sec. 4.2) and we plotted the results in Fig. 1(b) (left). λ is shown on the x-axis and entropic value of state 1 is shown on the y-axis (Eq. 15). As expected the optimal policy is risk-seeking if the λ is positive, i.e. the policy chooses "bet, bet", and risk-aversive if the λ is negative, i.e. the policy chooses "no, no" at state 1 and 2.

The curve can intuitively be explained with the help of the mean-variance criterion. For small λ the entropic map approximates the mean-variance criterion and for the "no, no" policy both the variance is 0 and hence we expect a value close to mean which is 0.92.

The second map, *mixed entropic map*, is constructed from entropic maps,

$$R(v|x,a) := \gamma^{-1} \log \mathbb{E}^Q_{x,a}\left[e^{\gamma v}\right], \gamma = \begin{cases} \lambda & \text{if } \mathbb{E}^Q_{x,a}\left[e^{\lambda v}\right] > 1 \\ -\lambda & \text{otherwise} \end{cases}$$

for some $\lambda \geq 0$. It is easy to check that this map satisfies the axioms of prospect maps and is convex (risk-seeking) if $v \geq 0$ and concave (risk-aversive) if $v \leq 0$. However, in the whole space the measure is neither convex nor concave. The risk-preference is controlled by $\lambda \geq 0$. The result is shown in Fig. 1(b) (right). For small λ, the optimal policy chosen by this measure is as expected: risk-seeking "bet" when facing gain but risk-aversive "not bet" when facing loss.

Grid World: Q-Learning In this experiment, we consider the 11×11 grid world depicted in Fig. 2(a). The agent will obtain a small reward $r_S = 3$ if hitting the upper-right corner (marked "S") and a large reward $r_L = 15$ if arriving at the lower-left corner (marked "L"). The shadowed grids denote "dangerous" states where the agent will be punished by a negative reward $r_D = -5$. In real applications, "dangerous" states can model the uncertain areas where punishments might be incurred.

The agent has four actions, "left", "right", "up" and "down" at each state. Choosing "left", the agent will deterministically go to the left neighboring state. If the agent is on the left boundary of the grid world, choosing "left" the agent will stay at the current state. The transitions of other 3 actions are defined similarly. However, if the agent is in one "dangerous" state and chooses "left", the probability of arriving at the left neighboring state (probability of escape) is only $P_e < 1$ and the agent will stay at the"dangerous" state with probability $1 - P_e$. For all 4 actions, the probability of escape is set to the same value.

We start from the upper-left corner (marked by a black point). The classical MDP with high discount factor will choose the path hitting the large-reward state "L" (depicted with red arrows in Fig. 2(a)). However, since the large-reward state is surrounded by dangerous states, a risk-aversive agent will dislike the policy that generates the highest average reward and instead choose the safer path (black arrow path in Fig. 2(a)) that avoids the all dangerous states.

We apply Q-learning to solve optimization problems for both classical MDP and entropic map (cf. Sec. 6). The same setup for both maps is used. Totally 200 episodes and each episode 250 steps are run. At the beginning of each episode, the state s_t is reset to the start state (upper-left corner). The learning rate α of each state-action pair decays propositional to the times of visiting the pair. The action of each update step is chosen by a ϵ-greedy policy where ϵ decays propositional to episode number. Therefore, in early episodes, ϵ is high to encourage exploration. At the end of each episode, a greedy policy is calculated by the current Q-value and the performance is evaluated by the value of the start state, v_1, generated by the learned policy. The Q-learning is considered to have converged, if the value is very close to the optimal value of the start state v_1^* that is calculated by value iteration (we know the transition model and reward model) before running Q-learning algorithm.

Fig. 2(b) plots the absolute difference $|v_1 - v_1^*|$ during learning procedure. The black curve depicts the result averaged over 200 random trials of entropic map with $\lambda = 0.01$ and $\alpha = 0.9$ while the red dotted curve show the average result of classical MDP with$\lambda = 0.9$. Both maps have the same optimal policy that finally hits the large-reward state and have the similar optimal value $v_1^* \approx 36$. Therefore, the Q-learning for both maps should have similar behavior and performance, which is confirmed by the Fig. 2(b). They have almost the same decay speed and therefore the same convergence speed.

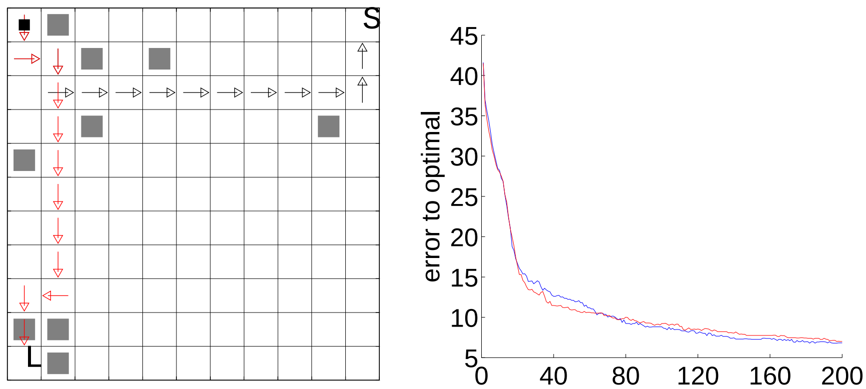
References

[1] B. Acciaio, H. Föllmer, and I. Penner. Risk assessment for uncertain cash flows: model ambiguity, discounting ambiguity, and the role of bubbles. *Arxiv preprint arXiv:1002.3627*, 2010.

[2] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[3] D.P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific Belmont, MA, 1995.

[4] D.P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*, volume 139. Academic Pr, 1978.



(a) 11 × 11 Grid World

(b) Error between learn value and optimal value

Figure 2: Grid World

[5] VS Borkar and SP Meyn. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, pages 192–209, 2002.

[6] A. Chateauneuf and M. Cohen. Cardinal extensions of the eu model based on the choquet integral. *Decision-making Process*, pages 401–433, 2008.

[7] K.J. Chung and M.J. Sobel. Discounted mdps: distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25:49, 1987.

[8] S.P. Coraluppi and S.I. Marcus. Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes. *Automatic Control, IEEE Transactions on*, 45(3):528–532, 2000.

[9] K. Detlefsen and G. Scandolo. Conditional and dynamic convex risk measures. *Finance and Stochastics*, 9(4):539–561, 2005.

[10] J.A. Filar, LCM Kallenberg, and H.M. Lee. Variance-penalized markov decision processes. *Mathematics of Operations Research*, pages 147–161, 1989.

[11] H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

[12] S. Gaubert and J. Gunawardena. The perron-frobenius theorem for homogeneous, monotone functions. *Transactions American Mathematical Society*, 356(12):4931–4950, 2004.

[13] C. Gollier. *The Economics of Risk and Time*. The MIT Press, 2004.

[14] A. Gosavi. A risk-sensitive approach to total productive maintenance. *Automatica*, 42(8):1321–1330, 2006.

[15] O. Hernández-Lerma. *Adaptive Markov Control Processes*, volume 79. Springer, 1989.

[16] O. Hernández-Lerma and J.B. Lasserre. *Discrete-time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[17] G.A. Holton. *Value-at-risk: theory and practice*. Academic Press, 2003.

[18] Y. Huang and LCM Kallenberg. On finding optimal policies for Markov decision chains: a unifying framework for mean-variance-tradeoffs. *Mathematics of operations research*, pages 434–448, 1994.

[19] G.N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, pages 257–280, 2005.

[20] A. Jobert and L.C.G. Rogers. Valuations and dynamic convex risk measures. *Mathematical Finance*, 18(1):1–22, 2008.

[21] D. Kahneman and A. Tversky. Prospect theory: an analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[22] H.J. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Verlag, 2003.

[23] S. Mannor and J. N. Tsitsiklis. Mean-variance optimization in markov decision processes. Submitted. `http://www.mit.edu/~jnt/Papers/P-10-mv-MDP-sub.pdf`, 2010.

[24] R.D. Nussbaum. *Hilberts Projective Metric and Iterated Nonlinear Maps*. American Mathematical Society, Providence, RI, 1988.

[25] W. Ogryczak and A. Ruszczynski. From stochastic dominance to mean-risk models: Semideviations as risk measures1. *European Journal of Operational Research*, 116(1):33–50, 1999.

[26] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., 1994.

[27] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[28] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, pages 1–27, 2010.

[29] L.J. Savage. *The foundations of statistics*. Dover Pubns, 1972.

[30] M.J. Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

[31] S.P.Coraluppi. *Optimal Control of Markov Decision Processes for Performance and Robustness*. PhD thesis, University of Maryland, 1997.

[32] C. Starmer. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38(2):332–382, 2000.

[33] R.S. Sutton and A.G. Barto. *Reinforcement learning*, volume 9. MIT Press, 1998.